Published online 12 October 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D381–D387***
doi:10.1093/nar/gkr847

Polbase: a repository of biochemical, genetic and structural information about DNA polymerases

Bradley W. Langhorst[1], William E. Jack[1], Linda Reha-Krantz[2] and Nicole M. Nichols[1,*]

[1]New England Biolabs, 240 County Road, Ipswich, MA, USA and [2]Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada

Received August 15, 2011; Revised September 19, 2011; Accepted September 21, 2011

ABSTRACT

Polbase (http://polbase.neb.com) is a freely accessible database of DNA polymerases and related references. It has been developed in a collaborative model with experts whose contributions reflect their varied backgrounds in genetics, structural biology and biochemistry. Polbase is designed to compile detailed results of polymerase experimentation, presenting them in a dynamic view to inform further research. After validation, results from references are displayed in context with relevant experimental details and are always traceable to their source publication. Polbase is connected to other resources, including PubMed, UniProt and the RCSB Protein Data Bank, to provide multi-faceted views of polymerase knowledge. In addition to a simple web interface, Polbase data is exposed for custom analysis by external software. With the contributions of many polymerase investigators, Polbase has become a powerful research tool covering most important aspects of polymerases, from sequence and structure to biochemistry.

INTRODUCTION

DNA Polymerases are responsible for faithful replication of DNA and maintenance of genomic integrity via repair and recombination. These enzymes catalyze the addition of free nucleotides to the 3′-end of a growing deoxyribonucleic acid polymer. The polymerase forms a complex with the template strand, priming strand and incoming nucleotide to catalyze the addition of a specific dNTP to the priming strand.

Since the discovery of the first DNA polymerase by Arthur Kornberg 60 years ago (1), a variety of specialized polymerase types with roles in the many aspects of genome replication and maintenance have been revealed (2).

Polymerase nomenclature has evolved over time to reflect the changing polymerase landscape (3,4). The current system (4) categorizes polymerases into seven families based on sequence similarity and biological roles. Recently, exploration of repair and *trans*-lesion polymerases in families X and Y (5,6) have added nuance to the balance between faithful replication and cell survival.

In addition to their biological functions, polymerases have proven to be pillars of biotechnology. The discovery of thermostable polymerases (7) enabled the Polymerase Chain Reaction (PCR), forever changing molecular biology (8). Specialized polymerases are at the core of the current generation of DNA sequencing platforms and a growing variety of diagnostic and detection technologies.

Polymerases play a role in various diseases, including many genetic disorders, viral infections and cancers. Thus, understanding basic properties of these enzymes has been crucial to diagnosis and treatment. Despite their central biological function and importance to biotechnology, no single repository of polymerase information existed before Polbase.

POLBASE IMPLEMENTATION

Polbase is a collaborative, open database focused exclusively on DNA polymerases. It is intended to provide a unified information resource for both polymerase experts and those just entering the field. Polbase does not attempt to replace existing protein and genetic information resources. Instead, Polbase compiles the information available in external resources, extending it with results extracted from the primary literature and polymerase-specific features. Since authors are in the best position to enter the important results from their publications quickly and accurately, we ask them to perform this critical function. Polbase was begun with the contribution of references and curation efforts from a small number of founding collaborators. More recently the larger polymerase community has been engaged to finish the work of cataloging the wealth of information in this mature field. By spreading the effort of maintaining Polbase throughout the polymerase community, minimal effort is required of any single research group.

*To whom correspondence should be addressed. Tel: +1 978 380 7266; Fax: +1 978 921 1350; Email: nichols@neb.com

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

New polymerase references are discovered and added to Polbase via automated tools or manual submission. As new papers are added, the corresponding author is contacted by email and asked to complete and validate the Polbase representation of their work.

Each reference follows a linear path as it is imported into Polbase. After a reference is created, its topics are specified by picking from a short list of polymerase relevant topics. Polymerases are then added to the reference's entry. If a new polymerase is encountered, a place-holder polymerase entry is created. After all polymerases and mutants are listed, the paper's primary results and the relevant experimental conditions are added by the contributor and indexed for searching. When all the results from a paper have been entered, an author 'validates' Polbase's representation of their work. Users may track the status of each reference in a queue on their personal account page. Polbase captures increasing amounts of information as the paper progresses through this pathway.

Topic and polymerase specifications allow Polbase to expand PubMed's author, abstract and title searching to include not only reference searches by polymerase (including aliases), polymerase family and host organism, but also by polymerase-specific topics. Topics such as 'Nucleotide Substitution' and 'Kinetic Parameters' improve searches for references and primary data (Table 1) and simplify results entry. Polbase also extracts polymerase features from the RCSB Protein Data Bank (PDB) (9) allowing users to find structures by polymerase, family and presence or absence of DNA in the protein structure.

Relevant experimental details from each paper are stored and displayed with results to allow users to assess information in context. Available contextual information varies by result type, and includes details such as salt concentration, presence of accessory factors, reactants, experimental technique used, etc. This system can be readily extended to accommodate new contexts as they arise or increase in importance to the user community. All results are also linked to their source publication so they can be easily found in their original context.

Polbase avoids editorial positions on the quality of any given result and does not present averaged values. Instead, summarized results are presented as ranges and link to individual results, which are presented with their context allowing the user to assess which are most relevant. Authenticated users and authors of primary data are encouraged to mark references as completely and/or correctly represented in Polbase to facilitate such assessment.

References are not limited to journal articles; negative results and other data not suitable for traditional publication can be made publicly available in Polbase with appropriate indications about the source of this information.

Polbase is built on a carefully designed table structure in a proven relational database system (10) so that the compiled information will be available to future applications in addition to its current web interface.

INFORMATION SOURCES

Polbase is tightly integrated with existing databases to provide a polymerase-centric perspective without unnecessary duplication of effort. All reference entries are linked to PubMed entries where they exist. New references in PubMed are discovered, imported and associated with polymerases by semi-automated tools. If new publications contain a corresponding author's email address or have an author with an active Polbase account, the paper is added to the author's 'queue'. Authors receive notification of new references according to their contact preferences.

Polbase is updated daily with all new reports of polymerase structures in the PDB (E.C.# 2.7.7.7 or 2.7.7.49). These structures are linked with existing polymerase entries where possible. Polymerase entries are linked with UniProt (11). Host organisms are linked with the NCBI Taxonomy tree (12).

WEBSITE ORGANIZATION

Browsing

Polymerases, references, structures and authors are available in list form via links on the main navigation bar at the

Table 1. Advantages of polymerase-specific search features

Compared to:	External search terms:	Polbase search terms:	Example searches possible in Polbase
Reference searching at PubMed	Title, Abstract, Author names, MeSH terms	Title, Abstract, Disambiguated Authors, Polymerase topics, Polymerase name, Polymerase family, Polymerase relationships, Organism, Polymerase property...	Reference covering family B polymerases and fidelity All references by polymerase author Shonen Yoshida (excluding another researcher of the same name) All of Arthur Kornberg's publications on DNA polymerase kinetics Summarized view of exonuclease activities of wild type and mutant phage T4 DNA polymerases
Structure searching at PDB	Name, Ligands, PDB ID	Polymerase, ±DNA in crystal, PDB ID, ±Mutants, Polymerase family...	All structures of family B polymerases with DNA in the crystal All structures of wild type and mutant phage RB69 DNA polymerases

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

left of every page (Figure 1). These indices can be filtered and sorted to quickly find a specific item of interest. Any list view can be sorted by column with a click on its header.

The polymerase index page (Figure 1) includes polymerase name and family with a summary of selected properties. Additional tabs provide structure information and connectivity with related polymerases (mutants, isoforms, etc.). The reference index presents high-level summary information about publications in Polbase including the number of references by year, polymerase family and journal. A browsable index of references in Polbase is also available. It allows searching and sorting by authors, titles and dates of publication. The authors index includes author's names and a count of how many of their publications are listed in Polbase. The structure index catalogs all polymerases and mutants with structure information. It includes a field indicating the presence or absence of DNA in the structure, the structure title and the polymerase it is linked to.

Sorting and filtering features on the top-level index pages facilitate exploration of Polbase content and location of polymerases having shared properties (Figure 1E and H).

Individual Pages

Individual pages are available for most categories of information stored in Polbase, including polymerases, references, structures, authors, etc.

A *Polymerase Page* (Figure 2) contains information about relationships with other polymerases (mutants, digestions, etc.), an interactive map of known mutations (with each dot associated with a mutant and linking to its own record and results), a list of relevant references, PDB structure information (if available), the host organism name, and a summary of results linked to the polymerase.

Each *Reference Page* indicates which polymerases (wild-type or mutant) are covered in the reference, the citation details including an abstract, Polbase import pipeline stage, results links (for each polymerase) and links to this paper in both PubMed and at the publisher's website (if available).

Each *Organism Page* displays the list of organisms and their kingdoms with the numbers of known polymerases (including mutants).

Structure Pages include PDBsum (13) and Protein Data Bank entry links, template and/or primer DNA information, and a Polbase polymerase entry link.

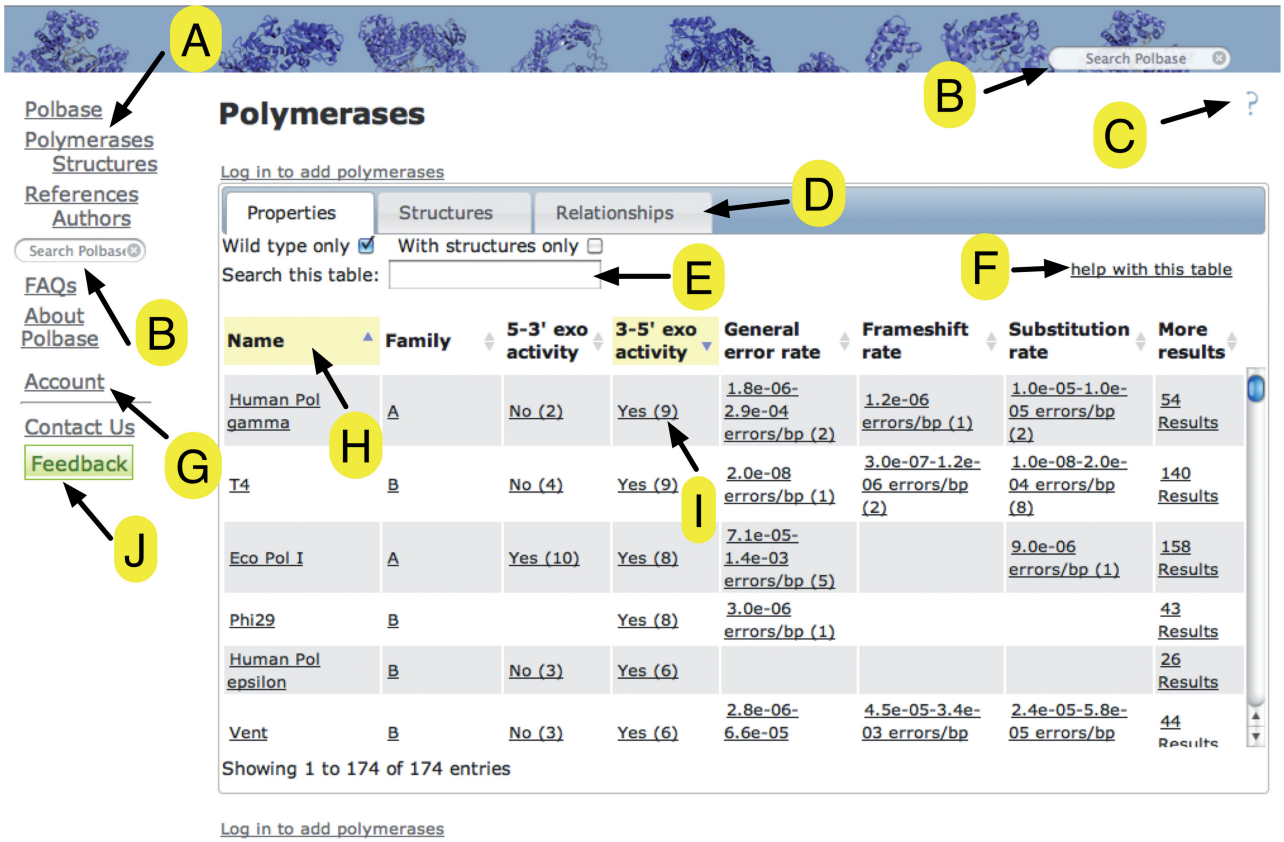

Figure 1. Polbase Navigation, Polymerase Index Page. A variety of features are available throughout Polbase, including (A) Links to Polbase features, (B) Polbase search features, (C) page-specific help, (D) tabs to categorize information, (E) selective filter tool to display only rows matching a search term, (F) more details about how to use Polbase tables, (G) link to a user account page with publications, polymerases, searches, etc., (H) sortable column titles, (I) numbers to indicate how many results contribute to summary values, (J) feedback link to aid community driven development.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

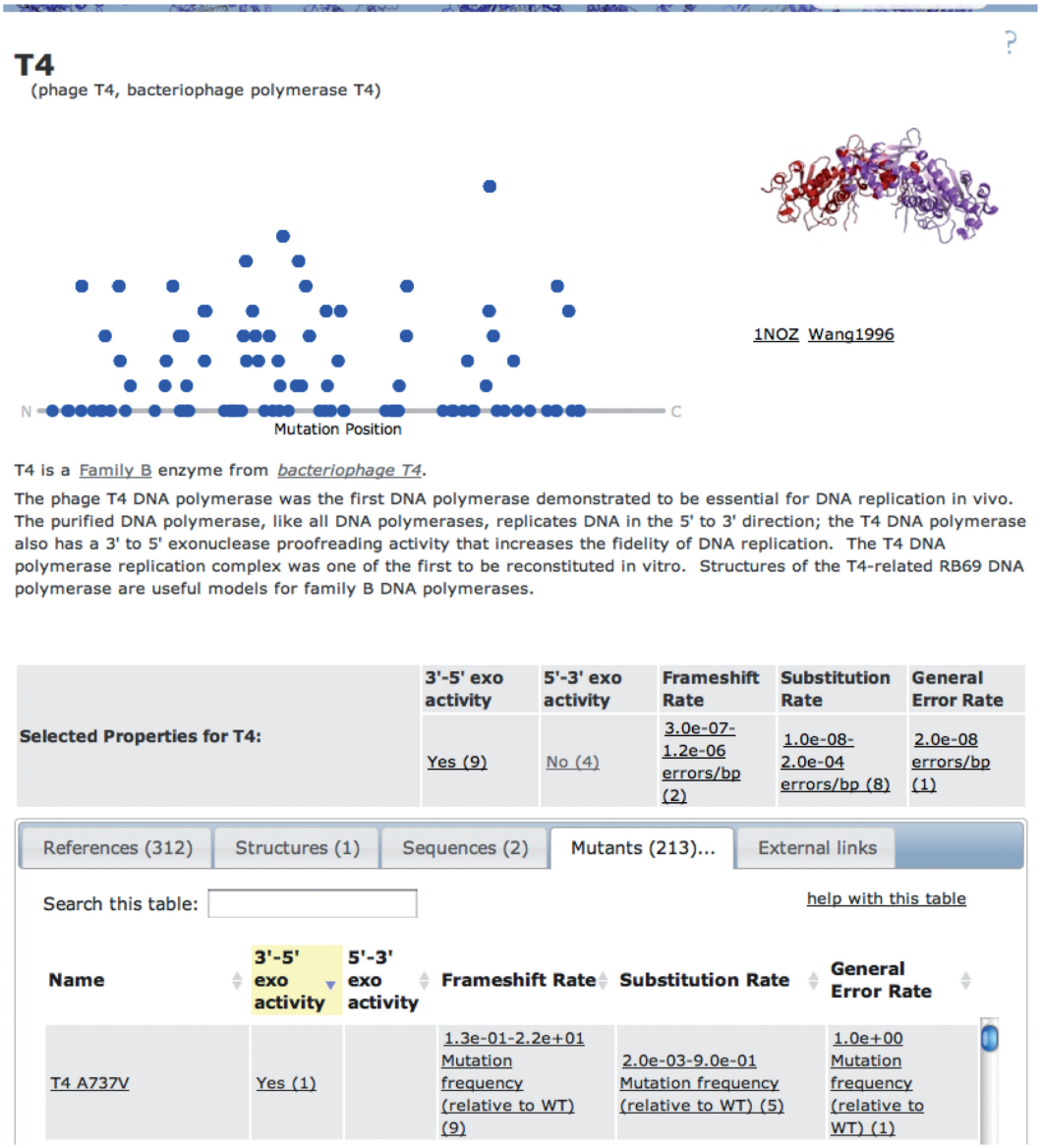

Figure 2. Example polymerase page. See text for description.

An *Author Page* (Figure 3) presents an author's publication history in graphical and tabular form. It also displays a bar graph indicating the number of an author's publications on each polymerase.

Author pages include all relevant publications by that author, and exclude publications from other authors with similar names. This simple requirement is complicated by the fact that authors may publish under multiple names, or share the same name. The identification problem is compounded by the many forms an author's name may take on (e.g. Lehman, I.R. Lehman I, I. Robert Lehman, etc.). Polbase uses authors' last names and first initials to construct lists of potentially matching authors. Matching authors are selectively merged using an iterative disambiguation algorithm that considers co-authorship, publication years and topics. Authors are manually split or merged in case of erroneous mergers or missed matches.

Search features

Polbase features a text search tool at the top of each page that provides a simple interface to the search index. In addition to the typically indexed publication fields (title, abstract, author names etc.), Polbase also maintains correlated indexes of authors, organisms, polymerase-specific properties and polymerases. A search for 'T4' produces the expected T4 polymerase record. In addition, correlated indices allow this search to return relevant authors, journal articles and the host organism, even when those items do not directly include the search term.

An advanced search tool allows a user to find all polymerases or references containing information about a specific topic, or search based on organism, family, etc (Figure 4). Because Polbase search is focused exclusively on DNA polymerases, search results are more relevant than comparable searches at PubMed, UniProt or PDB.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

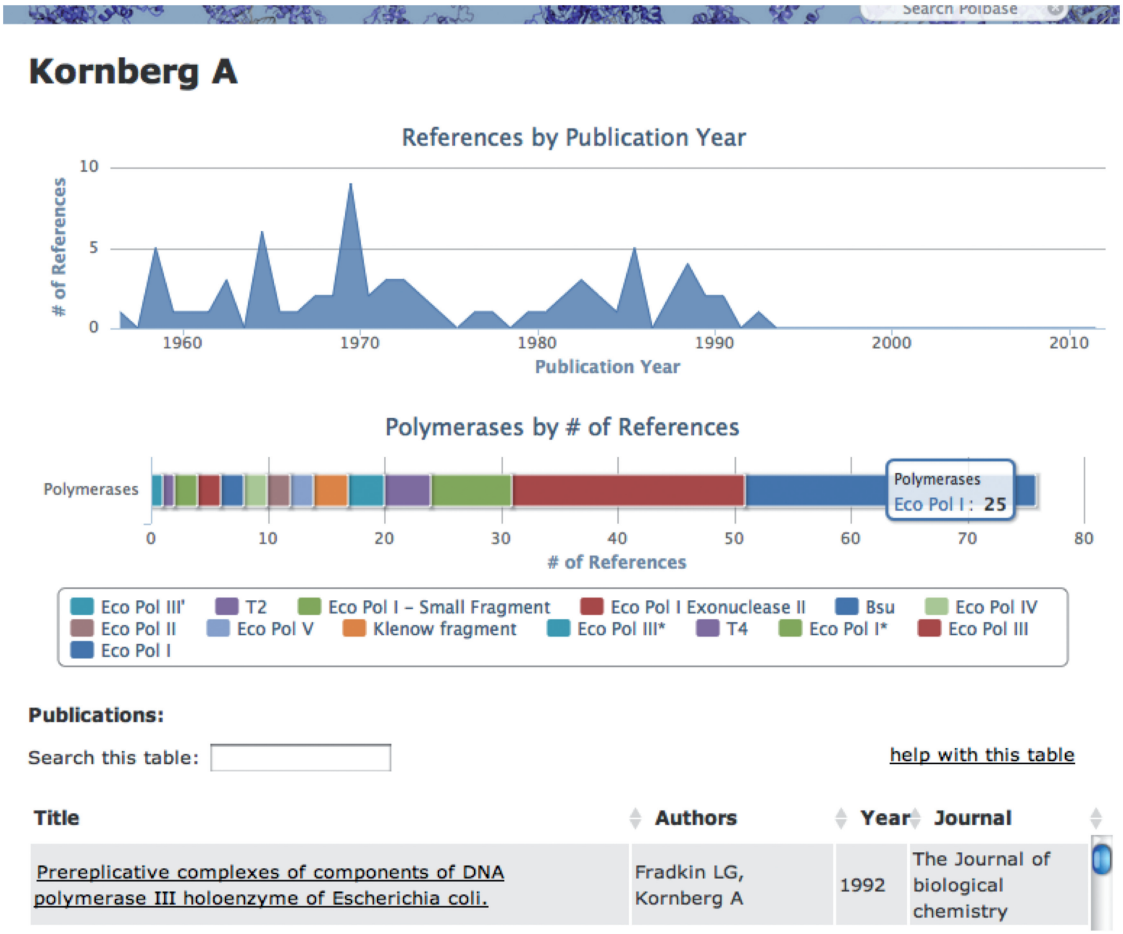

Figure 3. Example author page. See text for description.

Automated notifications

Users may 'watch' specific polymerases to receive updates at their chosen frequency. Likewise, searches may be saved and updates sent when results change or new results become available. These features can be managed using features on the account page (Figure 1G).

FUTURE WORK

Polbase expansion is user directed. There is a simple mechanism for users to suggest and vote on the priority of new features (Figure 1J).

With help from collaborators, Polbase's coverage of enzymes and specific polymerase activities continues to expand. Polbase continues to add sophistication to its automated reference discovery features. It currently uses a two-phase algorithm to identify papers for inclusion. First, a strict pass identifies papers that should certainly be included, skipping any ambiguous papers. The second pass uses frequently seen author names and permits inclusion of additional papers without incurring a high false positive rate. Polbase reference discovery features will continue to be improved, however it is unrealistic to expect that all relevant papers will be identified automatically. To assist with this process, polymerase experts are encouraged to contribute any missing polymerase-related publications. Polbase is able to accept entire reference libraries in most popular formats, instantly categorizing their contents according to a shared technique or polymerase and skipping any pre-existing references in the collection.

CONCLUSION

Polbase is the first open, on-line catalog of DNA polymerase information, it currently contains 183 wild-type and ~700 mutant polymerases in all 7 families spanning 102 organisms. Over 7300 references are currently indexed, covering 488 structures and more than 2900 discrete results and is constantly expanding. This open database provides a flexible, transparent resource to the diverse scientists who study and use polymerases in their work.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

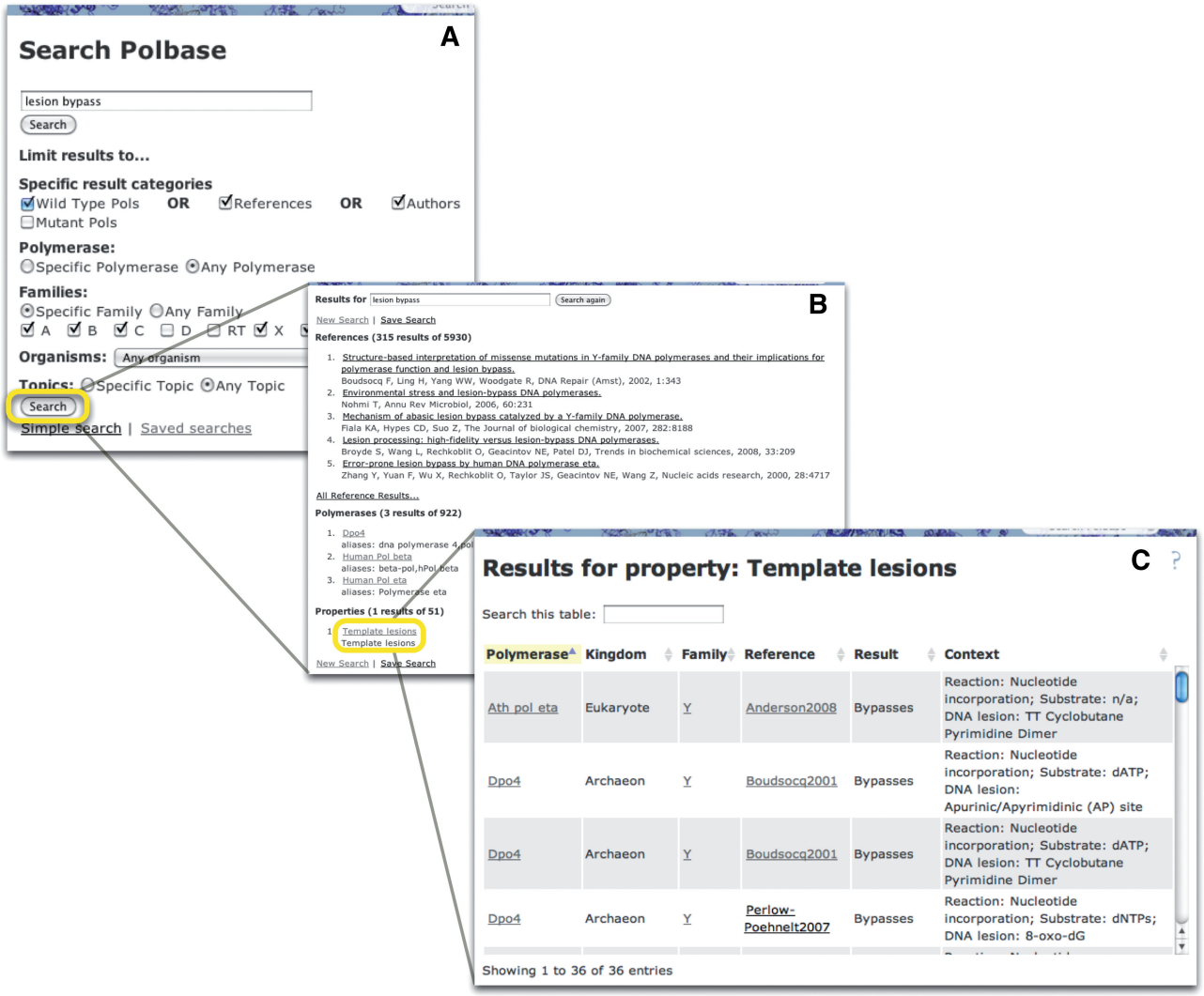

Figure 4. Polbase search features. Users can search Polbase for specific DNA polymerases, DNA polymerases in specific organisms, specific DNA polymerase families and a variety of DNA polymerase/DNA replication related topics. In this example, a user is searching for information about 'lesion bypass' (**A**) and identifies specific search terms in the Result categories, Polymerase, Families and Organisms fields. In (**B**), the user selects the relevant topic 'Template lesions' to reveal the detailed results displayed in (**C**).

AVAILABILITY

Polbase is a free, open, non-commercial resource, and its contents may be included (with attribution) in other software. Almost all resources are available in both human-readable HTML renderings as well as XML and JSON encodings for consumption by other software. URLs have been designed to be predictable and accessible. To request a document in XML format, simply append '.xml' to the URL (e.g. for all polymerases,/polymerases .xml).

ACKNOWLEDGEMENTS

The authors sincerely thank the following individuals for their contributions: Catherine Joyce for her early and continued help with the development of Polbase. William Beard, Edward Fox, Stuart Linn, Bruno Marchand, Stefan Sarafianos, Alexandra Vaisman, Samuel Wilson, Roger Woodgate, Lindsey Cantin, Devora Cohen-Karni, Eliot Chin, Adrienne Greenough and John Zou for database population, reference libraries, and valuable user feedback. Richard Roberts, Janos Posfai, Sanjay Kumar, Thomas Evans, Andrew Gardner, Jennifer Ong, Brendan Galvin, Ana Egaña, Julie Menin, Becky Kucera, Brendan Murphy and Paul Davis for valuable advice and planning assistance. Karen Otto, Ching-lun Lin, Thomas Peacock and Ruben Melo for infrastructure assistance, and Donald Comb and James Ellard for continued support for research at New England Biolabs.

Polbase would not be possible without the availability of PubMed and the RCSB PDB. Supporting software tools include: Ubuntu Linux, 10 Postgres, Memcache, Ruby, Rails, Sphinx and Thinking Sphinx, Apache and Phusion Passenger, UserVoice, Jquery, DataTables,

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

Authlogic, Declarative Authorization, Nokogiri, Paperclip, Formtastic, Will Paginate, BioRuby and Delayed Job.

FUNDING

New England Biolabs and a grant from the NIH SBIR program (#1R44GM087021). Funding for open access charge: New England Biolabs.

Conflict of interest statement. None declared.

REFERENCES

1. Bessman,M.J., Kornberg,A., Lehman,I.R. and Simms,E.S. (1956) Enzymic synthesis of deoxyribonucleic acid. *Biochim. Biophys. Acta*, **21**, 197–198.
2. Bebenek,K. and Kunkel,T.A. (2004) Functions of DNA polymerases. *Adv. Prot. Chem.*, **69**, 137–165.
3. Burgers,P.M., Bambara,R.A., Campbell,J.L., Chang,L.M., Downey,K.M., Hübscher,U., Lee,M.Y., Linn,S.M., So,A.G. and Spadari,S. (1990) Revised nomenclature for eukaryotic DNA polymerases. *Eur. J. Biochem./FEBS*, **191**, 617–618.
4. Burgers,P.M., Koonin,E.V., Bruford,E., Blanco,L., Burtis,K.C., Christman,M.F., Copeland,W.C., Friedberg,E.C., Hanaoka,F., Hinkle,D.C. *et al*. (2001) Eukaryotic DNA polymerases: proposal for a revised nomenclature. *J. Biol. Chem.*, **276**, 43487–43490.
5. Pata,J.D. (2010) Structural diversity of the Y-family DNA polymerases. *Biochim. Biophys. Acta*, **1804**, 1124–1135.
6. Yamtich,J. and Sweasy,J.B. (2010) DNA polymerase family X: function, structure, and cellular roles. *Biochim. Biophys. Acta*, **1804**, 1136–1150.
7. Chien,A., Edgar,D.B. and Trela,J.M. (1976) Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J. Bacteriol.*, **127**, 1550–1557.
8. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
9. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
10. Stonebraker,M. and Rowe,L.A. (1986) The design of POSTGRES. *SIGMOD Rec.*, **15**, 340–355.
11. UniProt Consortium,174. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
12. Wheeler,D.L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, 35D–40D.
13. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015